

観察研究における欠測データの統計解析： TARMOSガイドラインと新しいデータ解析・論文報告 のための枠組み

野間 久史
統計数理研究所
2024年10月9日
慶應義塾大学 第7回信濃町データサイエンスフォーラム
e-mail: noma@ism.ac.jp
URL: <http://www.ism.ac.jp/~noma/>

1

欠測データの統計解析

- ▶ 臨床研究・疫学研究において、計画されていた通りに、完全なデータが得られることはほとんどなく、なんらかの形でデータの欠測は、ほぼ確実に起こる
- ▶ 臨床試験においては、2010年頃に、欧米でガイドライン作成が開始され、国際医学誌でも、近年までに、その原理・原則がスタンダードなものとして採用されるようになってきた
- ▶ 一方、観察研究においては、最近まで、ガイドラインの整備はなされておらず、実践で用いられる方法の選択、また、論文上に報告される情報などの標準化はされていなかった

2

The STRATOS initiative

STRengthening Analytical Thinking for Observational Studies

STRATOS

INITIATIVE

Home

Topic Groups

Panels

Members

Links

Publications

News

Archive

Contact

About us

STRATOS initiative

The validity and practical utility of observational medical research depends critically on good study design, excellent data quality, appropriate statistical methods and accurate interpretation of results. Statistical methodology has seen substantial development in recent times. Unfortunately, many of these methodological developments are ignored in practice. Consequently, design and analysis of observational studies often exhibit serious weaknesses. The lack of guidance on vital practical issues discourages many applied researchers from using more sophisticated and possibly more appropriate methods when analyzing observational studies. Furthermore, many analyses are conducted by researchers with a relatively weak statistical background and limited experience in using statistical methodology and software. Consequently, even 'standard' analyses reported in the medical literature are often flawed, casting doubt on their results and conclusions. *An efficient way to help researchers to keep up with recent methodological developments is to develop guidance documents that are spread to the research community at large.*

These observations led to the initiation of the STRATOS (STRengthening Analytical Thinking for Observational Studies) initiative, a large collaboration of experts in many different areas of biostatistical research. *The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies. The guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests* (click to enlarge).

The Steering Group has decided to start with seven topics of general interest. Two topic groups were added later. Guidance documents will be developed by separate topic groups (TGs), each comprising experts in different area of statistical methodology, alongside applied researchers who may represent future end-users of the STRATOS documents. Each TG will start by developing guidance aimed primarily at level 2 statistical knowledge, which is perhaps slightly below state of the art. STRATOS structure and the initial [road map](#) (click to enlarge) from Sauerbrei et al 2014. The STRATOS initiative is closely connected to the [International Society of Clinical Biostatistics \(ISCB\)](#) and was launched at a half-day Mini-Symposium on the last day of the ISCB meeting in Munich, in August 2013.

Members Only Button

Members only

News

STRATOS Newsletter (May 2018)

IRS 2021

Invited Session

September 20, 2021

Milan, Italy

<https://stratos-initiative.org/>

3

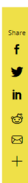
Topic Groups

1	MISSING DATA Chairs: James Carpenter, Kate Lee	2	SELECTION OF VARIABLES AND FUNCTIONAL FORMS IN MULTIVARIABLE ANALYSIS Chairs: Georg Heinze, Aris Perperoglou, Willi Sauerbrei	3	INITIAL DATA ANALYSIS Chairs: Marianne Huebner, Carsten Oliver Schmidt
4	MEASUREMENT ERROR AND MISCLASSIFICATION Chairs: Victor Kipnis, Pamela Shaw	5	STUDY DESIGN Chairs: Suzanne Cadarette, Mitchell Gail	6	EVALUATING DIAGNOSTIC TESTS AND PREDICTION MODELS Chairs: Ewout Steyerberg, Ben van Calster
7	CAUSAL INFERENCE Chairs: Els Goetghebeur, Ingeborg Waernbaum	8	SURVIVAL ANALYSIS Chairs: Michal Abrahamowicz, Malka Gorfine, Terry Therneau	9	HIGH-DIMENSIONAL DATA Chairs: Riccardo De Bin, Lisa McShane, Joerg Rahnenfuehrer

<https://stratos-initiative.org/>

4

STRATOS



The Missing Data Topic Group is one of the 9 topic groups within the [STrengthening Analytical Thinking for Observational Studies \(STRATOS\) Initiative](#). The objective of STRATOS is to provide accessible and accurate guidance in the design and analysis of observational studies. The guidance is intended for applied statisticians and other data analysts with varying levels of statistical education, experience and interests.

Aims of the Missing Data Topic Group

Members

TG1 represents a collaboration of 10 experts in missing data:

- James Carpenter (co-chair), London School of Hygiene and Tropical Medicine, London, UK
- Katherine Lee (co-chair), Murdoch Children's Research Institute and University of Melbourne, Melbourne, Australia
- Melanie Bell, University of Arizona, U
- Els Goetghebeur, Ghent University, Belgium
- Joseph Hogan, Brown University, US
- Rod Little, University of Michigan, US
- Andrea Rotnitzky, Harvard University, US
- Kate Tilling, University of Bristol, Bristol, UK
- Rosie Cornish, University of Bristol, Bristol, UK
- Rheanna Mainzer, Murdoch Children's Research Institute, Melbourne, Australia

Contact us

We are grateful to (in time order) [ESRC Research Methods Programme](#), the [ESRC Researcher Development Initiative](#), [ESRC Research Fellowship RES-063-27-0257](#), [ESRC Follow-On Funding scheme \(grant RES-189-25-0103\)](#), [MRC grant G0900724](#) and [MRC fellowship MR/K02180X/1](#) for supporting the development and maintenance of this site.

This site was formerly [missingdata.org.uk](#)

<https://www.lshtm.ac.uk/research/centres-projects-groups/missing-data#stratos>

5

JCE
JOURNAL OF CLINICAL EPIDEMIOLOGY

Submit Log in Register Sub

ORIGINAL ARTICLE | VOLUME 134, P79-88, JUNE 01, 2021

Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework

Katherine J. Lee ¹, A. ², ³ • Kate M. Tilling • Rosie P. Cornish • ... Joseph W. Hogan • James R. Carpenter •
on behalf of the STRATOS initiative • Show all authors

Open Access • Published: February 01, 2021 • DOI: <https://doi.org/10.1016/j.jclinepi.2021.01.008> •
Check for updates

Highlights
Abstract
Keywords
1. Background
2. Case study: The avon longitudinal study of parents and children
3. The framework
4. Discussion

Highlights

- Missing data are ubiquitous in medical research.
- Guidance is available, but missing data are still often not handled appropriately.
- We present a framework for handling and reporting analyses of incomplete data.
- This framework encourages researchers to think systematically about missing data.
- Adoption of this framework will increase the reproducibility of research findings.
- This article provides a much needed framework for handling and reporting the analysis of incomplete data in observational studies.
- The framework puts a strong emphasis on preplanning the statistical analysis and encourages transparency when reporting the results of a study.
- Adoption of this framework will increase the confidence in and reproducibility of research findings.

[https://www.jclinepi.com/article/S0895-4356\(21\)00010-X/fulltext](https://www.jclinepi.com/article/S0895-4356(21)00010-X/fulltext)

6

TARMOS framework

- ▶ Treatment And Reporting of Missing data in Observational Studies (TARMOS) framework
- ▶ STRATOS Initiative の Topic Group 1 (Missing Data) が作成した観察研究のための欠測データの取り扱いと研究報告のためのガイドライン
- ▶ 将来的に、観察研究の論文報告における国際スタンダードとなる可能性があり、その原理と方法についての理解は重要

Lee et al. (2021) 7

TARMOSの枠組み：3つのステップ

- ▶ (1) 研究計画の段階で策定する統計解析計画に、欠測データの取り扱いを明示するべきである。では、どのように解析計画を策定するべきか？
- ▶ (2) 統計解析をどのように行うか？解析計画に策定された通りに行うべきである。しかし、データが出てからでないと、計画した方法に必要な仮定の妥当性はわからない、予想外の欠測が起こる可能性もある。どう対処すべきか？
- ▶ (3) 欠測データの取り扱い・統計解析の結果を、論文上でどのように報告するべきか？

Lee et al. (2021) 8

Step 1. 統計解析計画を策定する

- ▶ 主要な統計解析，副次的な統計解析についての解析計画は、観察研究でも、事前に策定することが重要である
- ▶ 欠測データをどのように扱うのかについても、その中に含まれるべきであるということが、TARMOS frameworkでは明記されている！
- ▶ 方法・モデル、どの変数で調整をするか、補完回数などまで、事前に設定することが望ましいとされている

Lee et al. (2021) 9

欠測のメカニズム

- ▶ MCAR (missing completely at random) : 欠測は、他のいかなる要因とも関連せず、完全にランダムに起こる
- ▶ MAR (missing at random) : 欠測は「ランダムに起こる」わけではないが、そのメカニズムは、観測されているデータによって完全に説明することができる
- ▶ MNAR (missing not at random) : MARの否定。つまり、欠測のメカニズムには、観測されていない要因が影響しており、観測されているデータによって完全に説明することはできない

10

代表的な解析方法

- ▶ 完全ケース解析 (complete-case analysis)
- ▶ 多重代入法 (multiple imputation)
- ▶ 重み付け解析 (weighted analysis)

11

完全ケース解析

- ▶ 解析に必要な変数の組のうち（例えば、ロジスティック回帰をするのであれば、その結果変数・説明変数の組のうち）、1つでも欠測がある対象者を、解析対象集団から除外して解析するという方法
- ▶ 除外されたあとのデータセットには欠測はないので、通常の回帰モデルなどでの解析を行うことが可能
- ▶ ほとんどすべての統計ソフトはデフォルトでこの操作をしてしまうので注意！！

12

事例データセット

X_1	X_2	X_3	X_4	X_5
0	NA	1.129	0.049	1.084
1	0.694	0.602	1.018	-1.240
NA	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	NA
NA	NA	-1.527	-1.459	1.084
1	-0.243	-1.488	-1.449	-1.132
...	...	...	...	...

NA (not available) が、「欠測」を表します

13

完全ケース解析

X_1	X_2	X_3	X_4	X_5
0	NA	1.129	0.049	1.084
1	0.694	0.602	1.018	-1.240
NA	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	NA
NA	NA	-1.527	-1.459	1.084
1	-0.243	-1.488	-1.449	-1.132
...	...	...	...	...

除外 →

除外 →

少なくとも 1 つの変数に欠測がある対象者を除外します

14

多重代入法 (Multiple Imputation)

- ▶ データセット中の欠測値に対して、適当な代入値を埋め込むことによって、擬似的な完全データを作成し、これによって、すべての利用可能なデータを用いたCox回帰などの解析を行うことができるようにする
- ▶ 欠測してしまったデータの不確実性を反映するため、複数の補完値を生成し、その補完後のデータの複数の解析結果を統合することによって、不確実性を反映した分析結果を与えることができる

Rubin (1987) 15

欠測値の代入 (Imputation) とは？

X ₁	X ₂	X ₃	X ₄	X ₅
0	-0.125	1.129	0.049	1.084
1	0.694	0.602	1.018	NA
0	-0.761	1.229	0.922	-0.343
0	-0.809	-1.464	1.089	0.870
1	0.327	-1.527	-1.459	NA
1	-0.243	-1.488	-1.449	-1.132
...	...	...	...	...

-1.240

字義通り、欠測値を
なんらかの値で埋める
操作のこと

1.084

16

多重代入法 (Multiple Imputation)

X ₁	X ₂	X ₃	X ₄	X ₅	
0	-0.125	1.129	0.049	1.084	-1.240
1	0.694	0.602	1.018	NA	-0.898
					⋮
0	-0.761	1.229	0.922	-0.343	-1.338
0	-0.809	-1.464	1.089	0.870	1.890
1	0.327	-1.527	-1.459	NA	0.911
					⋮
1	-0.243	-1.488	-1.449	-1.132	1.084
...	...	...	...	...	

1つではなく、
複数の補完値を
用いる！！

17

Example : WCGS cohort (後半の事例データ)

```

x age      bmi sbp chol smoke      nael d
1  1  44 25.12681 NA   NA      1 0.10795017 0
2  0  50 24.38779 138 145      1 0.08503904 0
3 <NA> NA 25.12681 108  NA      1 0.05685942 0
4  1  45 24.70547 142 239      1 0.08908341 0
5  1  43 21.63297 110 192      0 0.02710404 0
6  1  43 25.84751 120 261      1 0.09454780 0
7  0  39 25.85015 126 212      1 0.08257678 0
8 <NA> 47 23.73116 136 203      0 0.09820397 0
9  1  45 23.69597  NA 200      1 0.10224847 0
10 1  46 24.69350 146 214      1 0.10795017 0
11 <NA> 59 25.88195 172  NA <NA> 0.08986772 0
12 <NA> 39 26.74163 142 195      0 0.06687382 0
13  0  NA 21.54179 116  NA      1 0.10224847 0
14  0  44 18.77121 148 189      1 0.10795017 0
15 <NA> 51 26.65525 130 301      1 0.08257678 0
16  1  55 25.57023 146 250      0 0.01795665 0
17  0  54 20.73197 114 179      0 0.08452936 0
18  0  43 21.52428 120 261 <NA> 0.10795017 0
19  1  42 23.10889 110 237 <NA> 0.12968930 0
20 <NA> 42 22.73030 110 268 <NA> 0.08664442 0

```

NA (not available) が欠測を意味する表示です。

18

1組目の補完データを代入

	x	age	bmi	sbp	chol	smoke	nael	d
1	1	44	25.12681	114	236	1	0.10795017	0
2	0	50	24.38779	138	145	1	0.08503904	0
3	0	46	25.12681	108	227	1	0.05685942	0
4	1	45	24.70547	142	239	1	0.08908341	0
5	1	43	21.63297	110	192	0	0.02710404	0
6	1	43	25.84751	120	261	1	0.09454780	0
7	0	39	25.85015	126	212	1	0.08257678	0
8	1	47	23.73116	136	203	0	0.09820397	0
9	1	45	23.69597	114	200	1	0.10224847	0
10	1	46	24.69350	146	214	1	0.10795017	0
11	1	59	25.88195	172	282	1	0.08986772	0
12	1	39	26.74163	142	195	0	0.06687382	0
13	0	40	21.54179	116	213	1	0.10224847	0
14	0	44	18.77121	148	189	1	0.10795017	0
15	1	51	26.65525	130	301	1	0.08257678	0
16	1	55	25.57023	146	250	0	0.01795665	0
17	0	54	20.73197	114	179	0	0.08452936	0
18	0	43	21.52428	120	261	0	0.10795017	0
19	1	42	23.10889	110	237	0	0.12968930	0
20	0	42	22.73030	110	268	0	0.08664442	0

多重代入法で、欠測データを埋めた結果です。補完後のデータセットには、欠測はひとつもなく、すべての対象者のデータを使って、Cox回帰などの多変量解析を行うことができます。多重代入法は「複数パターンの補完データを生成する」ことによって、欠測したデータを予測する不確実性を反映する方法になるわけですが、これが1組目の補完データとなります。

19

2組目の補完データを代入

	x	age	bmi	sbp	chol	smoke	nael	d
1	1	44	25.12681	146	235	1	0.10795017	0
2	0	50	24.38779	138	145	1	0.08503904	0
3	0	50	25.12681	108	261	1	0.05685942	0
4	1	45	24.70547	142	239	1	0.08908341	0
5	1	43	21.63297	110	192	0	0.02710404	0
6	1	43	25.84751	120	261	1	0.09454780	0
7	0	39	25.85015	126	212	1	0.08257678	0
8	0	47	23.73116	136	203	0	0.09820397	0
9	1	45	23.69597	112	200	1	0.10224847	0
10	1	46	24.69350	146	214	1	0.10795017	0
11	1	59	25.88195	172	280	0	0.08986772	0
12	0	39	26.74163	142	195	0	0.06687382	0
13	0	42	21.54179	116	209	1	0.10224847	0
14	0	44	18.77121	148	189	1	0.10795017	0
15	1	51	26.65525	130	301	1	0.08257678	0
16	1	55	25.57023	146	250	0	0.01795665	0
17	0	54	20.73197	114	179	0	0.08452936	0
18	0	43	21.52428	120	261	1	0.10795017	0
19	1	42	23.10889	110	237	1	0.12968930	0
20	0	42	22.73030	110	268	1	0.08664442	0

2組目の補完データで埋めた結果となります。1組目の補完データと少しだけ（不確実性の分だけ）、補完値のデータは違うことがわかります。「どの程度の不確実性を考慮するか」は、数学的なアルゴリズムで、適当に調整されています。

20

50組目の補完データを代入

	x	age	bmi	sbp	chol	smoke	nael	d
1	1	44	25.12681	118	263	1	0.10795017	0
2	0	50	24.38779	138	145	1	0.08503904	0
3	0	47	25.12681	108	249	1	0.05685942	0
4	1	45	24.70547	142	239	1	0.08908341	0
5	1	43	21.63297	110	192	0	0.02710404	0
6	1	43	25.84751	120	261	1	0.09454780	0
7	0	39	25.85015	126	212	1	0.08257678	0
8	1	47	23.73116	136	203	0	0.09820397	0
9	1	45	23.69597	136	200	1	0.10224847	0
10	1	46	24.69350	146	214	1	0.10795017	0
11	1	59	25.88195	172	266	0	0.08986772	0
12	1	39	26.74163	142	195	0	0.06687382	0
13	0	52	21.54179	116	282	1	0.10224847	0
14	0	44	18.77121	148	189	1	0.10795017	0
15	0	51	26.65525	130	301	1	0.08257678	0
16	1	55	25.57023	146	250	0	0.01795665	0
17	0	54	20.73197	114	179	0	0.08452936	0
18	0	43	21.52428	120	261	0	0.10795017	0
19	1	42	23.10889	110	237	0	0.12968930	0
20	1	42	22.73030	110	268	1	0.08664442	0

50組目の補完データで埋めた結果となります。このように、複数組生成された補完データセットを、すべてCox回帰などの方法で解析して、それらの結果を統合し、最終的なハザード比の推定値や信頼区間を計算するというのが、多重代入法による分析手法です。

21

多重代入法：これだけ理解しておこう！

- ▶ 100パターンほどの補完値を適当な数理的方法・アルゴリズムによって生成する
- ▶ 補完値を欠測に代入した、100パターンの擬似的な完全データをそれぞれ解析し、100通りのCox回帰の結果を得る
- ▶ 適当な方法（Rubin's Rule）によって、100通りのCox回帰の結果を統合し、最終的なハザード比の推定値と信頼区間を得ることができる
- ▶ 補完値（欠測データの予測）の不確実性を反映した最終的なハザード比の推定値と信頼区間を得ることができる！

22

重み付け解析

- ▶ 傾向スコア重み付け解析と同じ要領で、各対象者において、特定の変数 X が「観測される」「欠測する」という事象について、ロジスティック回帰を当てはめ、「欠測を起こさない確率（傾向スコアに相当）」を推定する
- ▶ 「欠測を起こさない確率」の逆数によって、各対象者を重みづけする解析（inverse probability weighting; IPW）
- ▶ 傾向スコアのIPTW法と同じ原理で、非ランダムな欠測メカニズムによって生じるバイアスの調整が可能

23

TARMOSによる解析方法についての大意①

- ▶ MARのもとでの解析方法としては、多重代入法（multiple imputation; MI）が主たる解析方法として言及がされている
- ▶ MIの実践上の利点
 - ▶ すべての主要な統計ソフトウェアに実装されている
 - ▶ 補助変数（auxiliary variable）の情報を組み込みやすい
 - ▶ 感度解析を実行しやすい
 - ▶ 大規模データの解析に使いやすい

TARMOSにおける補助変数とは、主たる解析モデルに含められる変数ではないが、MIなどの解析での欠測メカニズムや欠測データの分布の説明に有用な変数を指す。

Lee et al. (2021) 24

TARMOSによる解析方法についての大意②

- ▶ 重み付き解析（IPW法）は選択肢にはなり得るが、ガイドライン中で、そこまで主要な方法としての言及はされていない
 - ▶ 観察研究では、やはり、欠測確率のモデルの説明変数が欠測しているケースが大半（IPW法の前提として、欠測確率のモデルの説明変数には1つも欠測があってはいけない）
- ▶ 実践で用いるためにはまだ方法論・計算上の課題が多いとも言及されている（Lee et al., 2021）

25

TARMOSによる解析方法についての大意③

- ▶ 完全ケース解析は、MCARの仮定が妥当でない場合には推奨されないというのがこれまでの定説であった
- ▶ しかし、近年の研究によって、曝露・アウトカム・交絡要因に多くの情報が失われたわけであれば（例えば、欠測が5%未満の対象者にのみ起こった場合など）、どのように欠測を扱っても、ほとんど結果に大差はなく、完全ケース解析が許容されると言われるようになってきている（Madley-Dowd et al., 2019）

26

TARMOSによる解析方法についての大意④

- ▶ また、近年までの研究により、完全データ解析では、MCARの他にも、特殊な条件下でバイアスのない結果が得られることがわかっている
- ▶ 特に重要な条件：解析モデルにおけるすべての変数で起きた欠測の理由が、アウトカムと関連がないのであれば、曝露－アウトカムの関連は、完全ケース解析で、バイアスなく推定することができる（欠測は、MNARであってもOK！）
- ▶ MIによる解析では、この条件ではバイアスが入る
- ▶ 完全データ解析は、MIに対しての感度解析の1つとして使うことができる

Hughes et al. (2019), Lee et al. (2021) 27

TARMOSによる解析方法についての大意⑤

- ▶ MARの仮定が成り立たないと考えられるときには、MNARの仮定のもとでの感度解析を行うべき、とされている
- ▶ 感度解析に想定され得るシナリオは、細かい問題まで含めて議論をすると、無数に存在する
- ▶ 通常は、1つか2つ程度の特に重要な変数にフォーカスして分析を行うことが一般的である
- ▶ また、使用方法も、数理的に高度・複雑な方法はいくらかでもあるが、ごくシンプルな方法での評価を行うことも可能であると言及されている（後述）

28

Step 2a. データを調べる

- ▶ 欠測のメカニズムに関する分析を行う
- ▶ 1. 解析モデルに用いられるすべての変数について、欠測データの割合を示す表を示すこと。
- ▶ 2. 完全なデータが得られた人 vs. そうでない人、多くの変数に欠測があった人 vs. そうでない人 など、の間に観測された特徴についてまとめた表を示すこと。
- ▶ 3. 欠測メカニズムに関連する要因（予測因子）の評価を行うこと。完全データであるかどうかをアウトカム変数としたロジスティック回帰モデルに、関連し得る要因を説明変数としてモデル化して、分析を行う。

Lee et al. (2021) 29

Home Study at Bristol About Schools & faculties Research Business & partnerships News People

University of BRISTOL Avon Longitudinal Study of Parents and Children

Current students Current staff Alumni

search

Avon Longitudinal Study of Parents and Children

Participants
Researchers
COVID-19 research
News
A to Z index
About
Contacts
Media

"Our @30 clinic has launched. Our original babies are now turning 30, for the first time we are inviting all three generations to a single clinic, called @30."

Professor Nic Timpson, Principal Investigator

To mark Children of the 90s' 30th year, we have embarked on our largest data collection yet. We are inviting all three generations of participants to attend - the original Children of the 90s, their parents and children. [Read more](#)

Coronavirus (COVID-19)
Read about how we're contributing to vital COVID-19 research

Popular links
Complete your questionnaire
[Submit/amend a research proposal](#)
[View the latest data releases](#)
[Update your contact details](#)
[Download the interview schedule](#)

<http://www.bristol.ac.uk/alspac/> 30

Supplementary Table 2: Summary of the variables in the analysis model for the ALSPAC case study including the amount of data available for each variable, and a summary of the characteristics for the enrolled sample and those with complete records.^{1,2}

Characteristic ^{c,3}	^{c,3}	Available data (n=14,684) ^{c,4} N (%) ^{c,3}	Enrolled singletons and twins alive at one year and not withdrawn (n=14,684) ^{1,c,3}	Complete records ^{c,4} (n=3,313) ^{c,3}
Sex ^{c,3}	Male ^{c,4} Female ^{c,3}	14,684 (100%) ^{c,3}	7,536 (51%) ^{c,4} 7,148 ^{c,3}	1,559 (47%) ^{c,4} 1,754 ^{c,3}
Parity ^{c,3}	0 ^{c,4} 1 ^{c,4} 2+ ^{c,3}	12,924 (88%) ^{c,3}	5,770 (45%) ^{c,4} 4,539 (35%) ^{c,4} 2,615 (20%) ^{c,3}	1,628 (49%) ^{c,4} 1,181 (36%) ^{c,4} 504 (15%) ^{c,3}
Mother's education ^{c,3}	O level/lower ^{c,4} A level ^{c,4} Degree/higher ^{c,3}	12,412 (85%) ^{c,3}	8,022 (65%) ^{c,4} 2,791 (22%) ^{c,4} 1,599 (13%) ^{c,3}	1,800 (54%) ^{c,4} 932 (28%) ^{c,4} 581 (18%) ^{c,3}
Father's education ^{c,3}	O level/lower ^{c,4} A level ^{c,4} Degree/higher ^{c,3}	10,717 (73%) ^{c,3}	5,445 (51%) ^{c,4} 3,104 (29%) ^{c,4} 2,168 (20%) ^{c,3}	1,473 (44%) ^{c,4} 1,054 (32%) ^{c,4} 786 (24%) ^{c,3}
Mother's smoking ^{c,3}	Never smoked ^{c,4} Smoked, not in pregnancy ^{c,4} Smoking in pregnancy ^{c,3}	13,242 (90%) ^{c,3}	6,413 (48%) ^{c,4} 3,584 (27%) ^{c,4} 3,245 (25%) ^{c,3}	1,958 (59%) ^{c,4} 934 (28%) ^{c,4} 421 (13%) ^{c,3}
Paternal smoking (ever smoked) ^{c,3}	No ^{c,4} Yes ^{c,3}	10,690 (73%) ^{c,3}	4,419 (41%) ^{c,4} 6,271 ^{c,3}	1,624 (49%) ^{c,4} 1,689 ^{c,3}
Behavioural difficulties score at 81 months ^{c,3}	Median (IQR) ^{c,3}	7,289 (50%) ^{c,3}	6 (4-10) ^{c,3}	6 (4-9) ^{c,3}
Attainment score at 11 years ^{c,3}	Mean (SD) ^{c,3}	11,813 (80%) ^{c,3}	65% (16%) ^{c,3}	71% (14%) ^{c,3}
Smoking at 14 years ^{c,3}	No ^{c,4} Yes ^{c,3}	7,211 (49%) ^{c,3}	6,762 (94%) ^{c,4} 449 (6%) ^{c,3}	3,123 (94%) ^{c,4} 190 (6%) ^{c,3}
Outcome: attainment score ^{c,3}	Mean (SD) ^{c,3}	12,020 (82%) ^{c,3}	58% (18%) ^{c,3}	67% (13%) ^{c,3}

Note, there are 3,313 participants who have complete data on all of these variables required for analysis (23% of the original 14,684).^{c,4}

¹ Denominators vary because the variables come from different sources/questionnaires and have different completion rates.^{c,4}

Lee et al. (2021) 31

Supplementary Table 3: Predictors of being a complete case in the ALSPAC case study (n=14,684)^{1,c,4}

Characteristic ^{c,3}	^{c,3}	Crude odds ratio (95% confidence interval) ^{c,3}	Area under the curve ^{c,3}
Sex ^{c,3}	Male ^{c,4} Female ^{c,3}	1.00 ^{c,4} 1.25 (1.15, 1.35) ^{c,3}	0.53 ^{c,3}
Parity ^{c,3}	0 ^{c,4} 1 ^{c,4} 2+ ^{c,3}	1.00 ^{c,4} 0.89 (0.81, 0.98) ^{c,4} 0.61 (0.54, 0.68) ^{c,3}	0.54 ^{c,3}
Mother's education ^{c,3}	O level/lower ^{c,4} A level ^{c,4} Degree/higher ^{c,3}	1.00 ^{c,4} 1.73 (1.58, 1.90) ^{c,4} 1.97 (1.76, 2.21) ^{c,3}	0.57 ^{c,3}
Father's education ^{c,3}	O level/lower ^{c,4} A level ^{c,4} Degree/higher ^{c,3}	1.00 ^{c,4} 1.39 (1.26, 1.53) ^{c,4} 1.53 (1.38, 1.71) ^{c,3}	0.55 ^{c,3}
Mother's smoking ^{c,3}	Never smoked ^{c,4} Smoked, not in pregnancy ^{c,4} Smoking in pregnancy ^{c,3}	1.00 ^{c,4} 0.80 (0.73, 0.88) ^{c,4} 0.34 (0.30, 0.38) ^{c,3}	0.59 ^{c,3}
Paternal smoking (ever smoked) ^{c,3}	No ^{c,4} Yes ^{c,3}	1.00 ^{c,4} 0.63 (0.58, 0.69) ^{c,3}	0.56 ^{c,3}
Behavioural difficulties score at 81 months ^{c,3}	For each 1 point increase ^{c,3}	0.96 (0.95, 0.97) ^{c,3}	0.55 ^{c,3}
Attainment at 11 years ^{c,3}	For each 10% increase ^{c,3}	1.47 (1.43, 1.51) ^{c,3}	0.66 ^{c,3}
Smoking at 14 years ^{c,3}	No ^{c,4} Yes ^{c,3}	1.00 ^{c,4} 0.85 (0.70, 1.04) ^{c,3}	0.50 ^{c,3}
Outcome: attainment score ^{c,3}	For each 10% increase ^{c,3}	1.67 (1.61, 1.73) ^{c,3}	0.70 ^{c,3}

¹ Denominators in each analysis vary because the variables come from different sources/questionnaires and have different completion rates.^{c,4}

Lee et al. (2021) 32

ALSPAC studyでの解析事例

- ▶ ALSPAC studyでは、14,684人のうちの3,313人（23%）にしか、解析に必要なすべての変数において、完全なデータが揃っていなかった
- ▶ 完全データが得られた対象者は、第1子であったこと、女性であること、両親が高学歴であること、両親が非喫煙者であること、などの条件に合致する人が相対的に多かった
- ▶ 特に、アウトカムである学業成績と13歳時点での喫煙状況が、完全ケースであることと有意な関連があった
- ▶ Complete Case Analysisでバイアスが入らないとは考えにくい？

33

ALSPAC studyでの解析事例

- ▶ 以上のことから、(1) 完全ケースの数はそもそも相対的に少ない上に、(2) アウトカムがさまざまな変数の欠測に関連していると考えられる
- ▶ (2) から、完全ケース解析ではバイアスが生じる可能性がある
- ▶ また、(1) から、完全ケース解析では、除外される対象者が多く、推定精度を悪化させる可能性がある
- ▶ MIによる解析で、バイアスを低減し、精度を改善できることが期待される
- ▶ また、MNARでの欠測が起こっている可能性も否定できず、感度解析を行うことが重要であるとも考えられる

34

Step 2b. 解析計画に従って、解析を行う

- ▶ 以上の分析によって、解析計画に示された仮定が妥当なものであると考えられる場合には、事前計画された解析を行う
- ▶ 一方、ここまでのデータチェックなどのプロセスを経て、解析計画を変更する必要がある場合には、変更は認められる。ただし、その正当化（根拠の説明）が必要である

Lee et al. (2021) 35

Step 3. 論文上での報告について

- ▶ Methodsのセクションでは、解析において、欠測データがどのように扱われたかを示すべきである（感度解析も含め）
- ▶ そして、それが事前に設定されたもの（prespecified）であるか、あるいは、変更が加えられたものであるかについても報告がなされるべきである

Lee et al. (2021) 36

Step 3. 論文上での報告について

- ▶ 解析の結果は、臨床的な解釈に、欠測データの影響を踏まえた解釈を行い、報告を行うべきである
- ▶ 感度解析を含め、すべての解析の結果が類似したものとなるのであれば、欠測データが結果に与える影響は大きくないと判断できる（頑健性の確認）
- ▶ 一方、解析方法によって、結果が食い違うようであれば、その違いについての説明を行うことが重要である

Lee et al. (2021) 37

ALSPAC studyでの報告の事例

Table 2. Analysis of the relationship between smoking at 14 years and educational attainment at 16 years

Method of analysis	Number of observations in the analysis	Regression coefficient (95% CI)	P	% Of missing smoking values imputed as "smokers"
Primary analysis: Multiple imputation	14,684	-10.8 (-12.2, -9.4)	<0.001	13.3
Complete records analysis	3,153	-7.9 (-9.1, -6.7)	<0.001	N/A
Sensitivity analysis—sensitivity parameter = 0.1	14,684	-10.9 (-12.4, -9.4)	<0.001	14.2
Sensitivity analysis—sensitivity parameter = 0.25	14,684	-11.0 (-12.3, -9.6)	<0.001	15.5
Sensitivity Analysis — sensitivity parameter = 0.5	14,684	-11.0 (-12.3, -9.6)	<0.001	18.1
Sensitivity analysis—sensitivity parameter = 1	14,684	-10.7 (-11.8, -9.6)	<0.001	24.2
Sensitivity analysis—sensitivity parameter = 10	14,684	-4.3 (-4.7, -3.8)	<0.001	99.8

感度解析には、MIに基づくパターン混合モデルが用いられている（後述）

Lee et al. (2021) 38

Rで実際に解析をしてみよう

Article

August 25, 1975

Coronary Heart Disease in the Western Collaborative Group Study Final Follow-up Experience of 8 1/2 Years

Ray H. Rosenman, MD; Richard J. Brand, PhD; C. David Jenkins, PhD; et al.

> Author Affiliations

JAMA. 1975;233(8):872-877. doi:10.1001/jama.1975.03260080034016



Abstract

Clinical coronary heart disease (CHD) occurred in 257 subjects during eight to nine years of follow-up (average, 8 1/2 years) in a prospective study of 39- to 59-year-old employed men. Incidence of CHD was significantly associated with parental CHD history, reported diabetes, schooling, smoking habits, overt behavior pattern, blood pressure, and serum levels of cholesterol, triglyceride, and β -lipoproteins. The type A behavior pattern was strongly related to the CHD incidence, and this association could not be explained by association of behavior pattern with any single predictive risk factor or with any combination of them.

(JAMA 233:872-877, 1975)

<https://jamanetwork.com/journals/jama/article-abstract/337636>

39

事例データ : cohort72.csv

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	D	1	1	44	0	25.12681			1	3128	0					
2		2	0	50	0	24.38779	138	145	1	2888	0					
3		3			0	25.12681	108		1	2068	0					
4		4	1	45	0	24.70547	142	239	1	2939	0					
5		5	1	43	0	21.63297	110	192	0	1075	0					
6		6	1	43	0	25.84751	120	251	1	3008	0					
7		7	0	39	0	25.85015	125	212	1	2858	0					
8		8		47	0	23.73116	136	203	0	3026	0					
9		9	1	45	0	23.69597		200	1	3060	0					
10		10	1	46	0	24.6935	146	214	1	3159	0					
11		11		59	1	25.88195	172			2965	0					
12		12		39	0	26.74163	142	195	0	2398	0					
13		13	0		0	21.54179	116		1	3048	0					
14		14	0	44	0	18.77121	148	189	1	3159	0					
15		15		51	0	26.65525	130	301	1	2860	0					
16		16	1	55	1	25.57023	146	250	0	820	0					
17		17	0	54	0	20.73197	114	179	0	2883	0					
18		18	0	43	0	21.52428	120	261		3148	0					
19		19	1	42	0	23.10889	110	237		3366	0					
20																

40

事例データ : cohort72.csv

- ▶ x: 性格のタイプ (Type A or B; 0 or 1 ; 主たる曝露変数)
- ▶ age: 年齢 (歳)
- ▶ bmi: BMI
- ▶ sbp: 収縮期血圧
- ▶ chol: 血中コレステロール
- ▶ smoke: 喫煙の有無
- ▶ t: 心血管イベント発生までの時間 (日)
- ▶ d: 心血管イベント発生の有無 (1: 有, 0: 無)

41

事例データ : cohort72.csv

- ▶ WCGSのコホートの中から、3000人をランダムに選定
- ▶ 適当な欠測メカニズムモデルを用いて、いくつかの変数に、人工的な欠測を生じさせています
- ▶ Cox回帰モデルを主たる解析モデルとして、心血管イベントの発症と性格のタイプ (Type A or B) の関連を、種々のリスク要因 (年齢, BMI, 血圧, コレステロール, 喫煙) を調整した上で評価してみよう!

42

TARMOSの指針に基づく解析例

- ▶ ① 対象となる集団における、欠測データの集計・記述を行おう
- ▶ ② 欠測メカニズムについての分析を行い、どのような手法を用いるのが妥当かを検討しよう
- ▶ ③ ②の分析の結果をもとに、完全データ分析, MI, 感度解析から妥当な方法を選定し、実際にデータの解析を試みよう

43

データセットの読み込み

```
> setwd("D:\\¥juntendo2023")
>
> wcfgs <- read.csv(file="cohort72.csv", header=TRUE) # データセットの読み込み
> head(wcfgs, 20)
```

	ID	x	age	bmi	sbp	chol	smoke	t	d
1	1	1	44	25.12681	NA	NA	1	3128	0
2	2	0	50	24.38779	138	145	1	2888	0
3	3	NA	NA	25.12681	108	NA	1	2068	0
4	4	1	45	24.70547	142	239	1	2939	0
5	5	1	43	21.63297	110	192	0	1075	0
6	6	1	43	25.84751	120	261	1	3008	0
7	7	0	39	25.85015	126	212	1	2858	0
8	8	NA	47	23.73116	136	203	0	3026	0
9	9	1	45	23.69597	NA	200	1	3060	0
10	10	1	46	24.69350	146	214	1	3159	0
11	11	NA	59	25.88195	172	NA	NA	2965	0
12	12	NA	39	26.74163	142	195	0	2398	0

44

欠測データの集計

```
> library("gtsummary")
>
> wgs2 <- wgs %>% select(x, age, bmi, sbp, chol, smoke, t, d)
> # 集計を取りたい変数を選定
> table2 <- tbl_summary(wgs2)
>
> table2 # 要約統計量を一括計算し、出力します。
```

gtsummaryは、臨床研究の論文のTable 1などに提示する要約統計量を一括計算することができるパッケージです。

45

“gtsummary” の出力

インターネットブラウザなどに、右のような表データが出力されるかと思います。このデータを、要約統計量の集計結果としてそのまま使うことができます。

各変数で、欠測は「Unknown」と分類されています。

それぞれの要約統計量は、欠測データを除き、観測されたデータ（available data）から計算されています。カテゴリカル変数の割合も、観測されたデータの中での割合として計算されています。

Characteristic	N = 3,000 [†]
x	
0	1,286 (54%)
1	1,075 (46%)
Unknown	639
age	45 (42, 50)
Unknown	84
bmi	24.41 (22.98, 25.87)
Unknown	119
sbp	126 (118, 136)
Unknown	273
chol	221 (195, 250)
Unknown	396
smoke	
0	1,331 (57%)
1	1,005 (43%)
Unknown	664
t	2,940 (2,842, 3,038)
d	248 (8.3%)
[†] n (%); Median (IQR)	

46

欠測データの集計

```
> library("naniar")
>
> miss_var_summary(wcgs)

# A tibble: 9 × 3
  variable n_miss pct_miss
  <chr>     <int>     <dbl>
1 smoke      664      22.1
2 x          639      21.3
3 chol       396      13.2
4 sbp        273       9.1
5 bmi        119       3.97
6 age         84       2.8
7 ID          0        0
8 t           0        0
9 d           0        0
```

naniarは、欠測データの集計を行うための関数や、グラフィカルツールが含まれているパッケージです。

便利な機能は、他にもたくさんあるのですが、ここでは、欠測データの集計ツールのみ、ご紹介します。

miss_var_summary関数を使うと、データセット中にある、各変数の欠測の数と割合を、一括して計算してくれます。

47

欠測データの集計

```
> cc <- complete.cases(wcgs) # Complete-caseに対応する行番号を計算します。
> wcgs.c <- wcgs[cc,]
>
> wcgs3 <- wcgs.c %>% select(x, age, bmi, sbp, chol, smoke, t, d)
> # 集計を取りたい変数を選定
> table3 <- tbl_summary(wcgs3)
>
> table3 # Complete-caseの集団における、要約統計量を一括計算し、出力します。
```

complete.casesという関数で、データセット中の完全ケースに対応する対象者の行番号を一括計算することができます。
データセットを、対応する行に限定すれば、完全ケースのデータセットを作成することができます。

48

“gtsummary” の出力

インターネットブラウザなどに、右のような表データが出力されるかと思います。このデータを、要約統計量の集計結果としてそのまま使うことができます。

完全ケースに限定したデータセットなので、欠測はひとつもありません。対象者集団は、1544人に減っているため、半分ほどの対象者には、少なくとも1つの変数で欠測があったことになります。

Characteristic	N = 1,544 [†]
x	
0	956 (62%)
1	588 (38%)
age	45 (41, 49)
bmi	24.39 (22.83, 25.85)
sbp	124 (118, 134)
chol	214 (190, 239)
smoke	
0	949 (61%)
1	595 (39%)
t	2,952 (2,864, 3,047)
d	27 (1.7%)
[†] n (%); Median (IQR)	

49

要約統計量のまとめ

Characteristic		Available data (N=3000)	Summary for the observed data	Complete records (N=1544)
x (main factor)	1	2361 (78.7%)	1075 (45.5%)	588 (38.1%)
	0		1286 (54.5%)	956 (61.9%)
age		2916 (97.2%)	45 (42, 50)	45 (41, 49)
bmi		2881 (96.0%)	24.41 (22.98, 25.87)	24.39 (22.83, 25.85)
sbp		2727 (90.9%)	126 (118, 136)	124 (118, 134)
chol		2604 (86.8%)	221 (195, 250)	214 (190, 239)
smoke	1	2336 (77.9%)	1005 (43.0%)	595 (38.5%)
	0		1331 (57.0%)	949 (61.5%)
t (outcome)		3000 (100%)	2940 (2842, 3038)	2952 (2864, 3047)
d (outcome)	1	3000 (100%)	248 (8.3%)	27 (1.7%)
	0		2752 (91.7%)	1517 (98.3%)

※ 連続変数は、中央値とIQRを要約統計量に用いています。

50

完全ケースと関連する要因の分析

```
> library("epiDisplay")
> cc <- complete.cases(wcgs)      # Complete-caseに対応する行番号を計算します。
> wcgs$R <- as.numeric(cc)        # Complete-caseであるかどうかを、{0, 1} の変数で定義します。
>
> # 完全ケースであるかどうかと、各共変量の周辺的な関連性を単変量ロジスティック回帰で分析します。
>
> logistic.display(glm(R ~ x, family=binomial(link="logit"), data=wcgs))

Logistic regression predicting R

                OR(95%CI)          P(Wald's test) P(LR-test)
x: 1 vs 0 0.42 (0.35, 0.5) < 0.001          < 0.001

Log-likelihood = -1472.7338
No. of observations = 2361
AIC value = 2949.4676
```

51

完全ケースと関連する要因の分析

Characteristic		Odds-ratio (95%CI)	P-value
x (main factor)	1	0.42 (0.35, 0.50)	< 0.001
	0	1.00	
age		0.96 (0.95, 0.97)	< 0.001
bmi		0.95 (0.93, 0.98)	0.001
sbp		0.98 (0.97, 0.98)	< 0.001
chol		0.99 (0.99, 0.99)	< 0.001
smoke	1	0.58 (0.49, 0.69)	< 0.001
	0	1.00	
t (outcome)		1.00 (1.00, 1.00)	< 0.001
d (outcome)	1	0.10 (0.07, 0.15)	< 0.001
	0	1.00	

52

完全ケースと関連する要因の分析

- ▶ 考慮されているすべての変数が、有意な関連を示した
- ▶ 特に、アウトカムと欠測メカニズムは強い関連があることが示唆されている
- ▶ 完全ケース解析による分析からはバイアスが入ってしまうリスクがある！
- ▶ また、完全ケース解析では、半数ほどの対象者が除外されてしまうため、推定精度や検出力の損失が大きい
- ▶ 欠測メカニズムを考慮した多重代入法などの解析方法がより妥当であると考えられる

53

完全ケース解析

```
> library("survival")
>
> gml <- coxph(Surv(t,d) ~ x + age + bmi + sbp + chol + smoke, data=wcgs)
> summary(gml)
```

```
n= 1544, number of events= 27
(1456 observations deleted due to missingness)
```

	exp(coef)	exp(-coef)	lower .95	upper .95
x1	1.3063	0.7655	0.6095	2.800
age	1.0711	0.9336	1.0019	1.145
bmi	1.1269	0.8874	0.9597	1.323
sbp	0.9824	1.0179	0.9513	1.014
chol	1.0028	0.9973	0.9928	1.013
smoke1	1.2140	0.8237	0.5578	2.642

1456人に、少なくとも1つの説明変数に欠測があり、解析から除外されています。
既知の強いリスク要因のいくつかは、関連なしという結果になっています。

54

多重代入法

```
> library("mice")
> B <- 100      # number of imputations; in practice, B should be set to >= 100.
>
> attach(wcgs)
>
> nael <- nelsonaalen(wcgs, t, d)  # Nelson-Aalen estimator for cumulative hazard
function
> wcgs4 <- data.frame(x, age, bmi, sbp, chol, smoke, nael, d)
# ここに入れた変数の組で、連鎖方程式を組んで、MIを行う；Cox回帰の場合は、イベントの指示変数（d）と、
Nelson-Aalen推定値（nael）を必ず入れましょう！
>
> detach(wcgs)
>
> imp4 <- mice(wcgs4, m=B)      # Multiple imputation by MICE
```

Rのパッケージ mice で、MICE (multiple imputation by chained equation) による多重代入法を、簡単なコマンドで実行することができます。

55

上級者向け：MICEによる解析を行う上での注意

- ▶ 基本的には、Chained Equationには、解析モデルのアウトカム変数は入れましょう
- ▶ アウトカムが、モデル化される曝露・交絡変数のいずれかと関連がある場合（通常はそれを想定している）、アウトカム変数が入っていないとバイアスが入ってしまう
- ▶ 連続・2値変数は、そのまま入れればよい
- ▶ 生存時間の場合は、イベント指示変数とNelson-Aalen推定値（nelsonaalen関数で計算可）を入れると良い近似法によることが示されている（White and Royston, 2009）

56

MICEによる多重代入法

```
> ima4 <- with(imp4, coxph(Surv(t4,d) ~ x + age + bmi + sbp + chol + smoke))
> est4 <- pool(ima4)
> summary(est4)
```

	term	estimate	std.error	statistic	df	p.value
1	x1	0.579260092	0.178328975	3.248267	103.79639	1.564692e-03
2	age	0.057796930	0.012000618	4.816163	204.07706	2.848453e-06
3	bmi	0.063551504	0.026557025	2.393020	191.41599	1.767705e-02
4	sbp	0.016796705	0.004529328	3.708432	159.65890	2.873406e-04
5	chol	0.007190536	0.001937923	3.710433	116.33398	3.188196e-04
6	smoke1	0.583376552	0.187925209	3.104302	88.11244	2.564576e-03

with関数とpool関数で、多重代入法による補完後の統合解析の結果を出力することができます。結果は、回帰係数のスケールでしか出てこないの、ロジスティック回帰やCox回帰での分析を行う場合には、exp関数で変換をして、オッズ比やハザード比に直してから、結果の解釈をすることにしましょう。

57

MICEによる多重代入法

```
> sum.coxph(est4) # MICEによるCox回帰の結果を、ハザード比に変換します
```

	term	HR	95%CL	95%CU	P-value
1	x1	1.784717	1.253098	2.541873	1.564692e-03
2	age	1.059500	1.034725	1.084868	2.848453e-06
3	bmi	1.065614	1.011232	1.122921	1.767705e-02
4	sbp	1.016939	1.007882	1.026076	2.873406e-04
5	chol	1.007216	1.003358	1.011090	3.188196e-04
6	smoke1	1.792079	1.233581	2.603435	2.564576e-03

sum.coxphという関数は、自作の関数ですが、miceの出力から、ハザード比などを計算する関数となっています。3000人全員のデータを活用した解析ができています。年齢、BMI、血圧、コレステロール、喫煙などの既知のリスク要因は、明確に、心血管疾患の発症との関連を示しています。性格のタイプ（Type A or B）も、心血管疾患の発症と有意な関連を示しています。

58

感度解析 (1) : 最も極端なケースでの解析

- ▶ 主たる曝露変数に欠測を起こした対象者の曝露変数を、全員「曝露あり」「曝露なし」に分類した場合の解析
- ▶ あらゆる欠測メカニズムを考慮した上での最も極端なケースであり、それぞれのケースでの曝露効果の推定結果が、結論を変えない範囲に存在するのであれば、主たる解析の結果の頑健性が支持されたと考えることができる
- ▶ 一定水準以上、曝露変数の欠測割合が小さい条件で有効な方法（欠測が多いと、一般的に範囲は広く出すぎてしまう）
- ▶ やや難しいですが、プログラム5をご参照ください

59

感度分析の結果

	HR	95%CI		P-value
多重代入法	1.78	1.25	2.54	< 0.001
曝露の欠測をすべて「曝露あり」にした感度解析	3.38	2.41	4.75	< 0.001
曝露の欠測をすべて「曝露なし」にした感度解析	0.66	0.50	0.88	< 0.001

x についての欠測が多いためか（21.3%）、HRの動く範囲は大きく、この極端な感度解析からは、曝露効果の推測についての頑健性を示すことはできないといえる。

60

感度解析 (2) : パターン混合モデル

- ▶ MICEのアルゴリズムに少し手を加えることで、MNARを想定したパターン混合モデルによる解析を行うことができる
- ▶ 関心のある変数 (e.g., 主たる曝露変数) の生成の際に、生成モデルの線形スコア (1次関数形式の予測スコア) に、一定の定数 δ を加える
- ▶ これによって、「観測されたデータ」と「欠測したデータ」の間で、データ生成分布が異なるというパターン混合モデルに基づく補完値の生成を行うことができる
- ▶ これも難しいですが、プログラム6をご参照ください

Lee et al. (2021), Yuan (2014) 61

パターン混合モデルによる感度解析の結果

	HR	95%CI		P-value
多重代入法	1.78	1.25	2.54	< 0.001
PMM ($\delta = -0.50$)	1.53	1.04	2.24	0.031
PMM ($\delta = -0.25$)	1.65	1.17	2.33	< 0.001
PMM ($\delta = -0.10$)	1.73	1.17	2.55	< 0.001
PMM ($\delta = 0.10$)	1.81	1.26	2.61	< 0.001
PMM ($\delta = 0.25$)	1.98	1.35	2.90	< 0.001
PMM ($\delta = 0.50$)	2.06	1.45	2.92	< 0.001

パターン混合モデルによって、xに欠測を起こした対象者における曝露確率の対数オッズが、-0.5から0.5の範囲に動かされても、効果の方向は一貫しており、一貫して有意差が出ることを確認することができた。この範囲内では、MARの仮定が崩れたもとでも、結果の頑健性を確認することができたといえるだろう。

62

感度解析の解釈

- ▶ 今回は、感度解析 (1) では、曝露効果の推測についての頑健性は示すことができなかったが、感度解析 (2) によって、相応のMARからの仮定の乖離があるもとでも、結論が頑健であることを示すことができた
- ▶ 主たる曝露変数の欠測割合が大きいと、(1) の方法はうまくいかないが、欠測割合が小さいと、任意のメカニズム（MNARを含む）のもとでの結論の頑健性を示すことができる強力な方法になる
- ▶ 例えば、同じ欠測メカニズムのもとで、x の欠測割合が5%程度になるように設定を変えた場合には...

63

曝露変数の欠測を 5% とした場合の解析結果

	HR	95%CI		P-value
多重代入法	1.82	1.35	2.44	< 0.001
完全ケース解析	1.59	0.63	0.90	0.112
曝露の欠測をすべて「曝露あり」にした感度解析	2.16	1.63	2.85	< 0.001
曝露の欠測をすべて「曝露なし」にした感度解析	1.27	0.98	1.63	0.068

HRの動く範囲は、大きく狭まり、任意の曝露変数の欠測パターンのもとで、アウトカムとの関連があることを示唆している。

64

曝露変数の欠測を 5% とした場合の解析結果

	HR	95%CI		P-value
多重代入法	1.82	1.35	2.44	< 0.001
PMM ($\delta = -0.50$)	1.70	1.27	2.28	< 0.001
PMM ($\delta = -0.25$)	1.74	1.30	2.32	< 0.001
PMM ($\delta = -0.10$)	1.80	1.33	2.42	< 0.001
PMM ($\delta = 0.10$)	1.81	1.36	2.43	< 0.001
PMM ($\delta = 0.25$)	1.86	1.39	2.49	< 0.001
PMM ($\delta = 0.50$)	1.91	1.43	2.54	< 0.001

同様に、パターン混合モデルによって、xに欠測を起こした対象者における曝露確率の対数オッズが、-0.5から0.5の範囲に動かされても、ほとんど結果は変わらないことが示されている。この分析の結果からも、主要な変数の欠測は小さく留めれば、不確実性には（やはり）悩まなくて済むということがわかる。

65

おわりに

- ▶ 観察研究における欠測データの取り扱いの方法として、STRATOS Initiativeから、TARMOSガイドラインが公表された
- ▶ 現状では、JAMA, BMJなどの一流誌などでも、それほどこの内容を反映した研究論文は多くないようであるが、今後、少しずつ、その普及は進んでいくものと思われる
- ▶ 分析に要求される水準は、やや高度であるため、今後は、観察研究においても、研究計画段階から、十分な技術を持つ統計家との協同が必要になってくるものと予想される

66

文献

- ▶ Hughes, R. A., Heron, J., Sterne, J. A. C., and Tilling, K. (2019). Accounting for missing data in statistical analyses: multiple imputation is not always the answer. *Int J Epidemiol* 48, 1294-1304.
- ▶ Lee, K. J., Tilling, K. M., Cornish, R. P., et al. (2021). Framework for the treatment and reporting of missing data in observational studies: The Treatment And Reporting of Missing data in Observational Studies framework. *J Clin Epidemiol* 134, 79-88.
- ▶ Madley-Dowd, P., Hughes, R., Tilling, K., and Heron, J. (2019). The proportion of missing data should not be used to guide decisions on multiple imputation. *J Clin Epidemiol* 110, 63-73.
- ▶ Royston, P., and White, I. R. (2011). Multiple Imputation by Chained Equations (MICE): Implementation in Stata. *J Stat Softw* 45, Issue 4.

67

- ▶ Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley.
- ▶ White, I. R., and Royston, P. (2009). Imputing missing covariate values for the Cox model. *Stat Med* 28, 1982-1998.
- ▶ White, I. R., Royston, P., and Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat Med* 30, 377-399.
- ▶ Yuan, Y. Sensitivity analysis in multiple imputation for missing data. SAS Institute Inc; 2014. <https://support.sas.com/resources/papers/proceedings14/SAS270-2014.pdf> Accessed November 22, 2022.

68